ChinaXiv合作期刊
图书情报工作
第64卷 第3期 2020年2月

高校图书馆读者借阅趋势线性回归建模预测探析*

■ 王红[1] 袁小舒[2] 原小玲[3] 黄建国[4]
[1] 山西财经大学图书馆 太原 030006 [2] 山西财经大学信息学院 太原 030006
[3] 太原科技大学图书馆 太原 030024 [4] 太原大然科技有限责任公司 太原 030006

摘 要：[目的/意义]通过馆藏图书分类和流通数据,发现读者特征与馆藏流通之间的关联,建立关系模型,通过模型拟合与预测,探索读者与图书流通之间的隐含规律,为图书馆智慧管理提供技术与手段的支持。[方法/过程]采用聚类和相关分析技术,提取读者宏观可观测特征,建立读者特征与图书分类之间直接和间接的映射关系,进而建立读者特征与分类图书流通量的回归模型,并验证模型有效性和优化模型拟合优度。根据有效模型,探索图书馆流通趋势,并挖掘读者宏观特征层面下所隐含的知识建构本质与规律,以及对图书流通产生的影响程度。[结果/结论]具有代表读者社会角色要求的专业学习方向、代表读者间群体互动效应的入学批次、读者群体数量3个有关读者的分类特征,能够有效拟合和预测图书流通量。预测结果表明,模型准确率较高,能够作为有效工具,为图书馆开展知识服务提供可靠的技术支持。

关键词：高校图书馆 流通预测 数据挖掘 线性回归
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.03.007

1 引言

图书流通量是读者与馆藏互动的结果,是联结读者与馆藏的关键指标,是衡量图书馆馆藏建设和读者服务质量的核心要素。建立基于不同类别图书流通量的描述模型,预测馆藏流通趋势,不仅能够提高图书馆的服务质量,预见性指导图书馆开展相关工作,还能为进一步揭示馆藏流通内在运行规律提供有力的支撑。高校图书馆读者与馆藏互动频繁,虽然图书和读者周期性更新频率较快,但围绕学科和教学服务的高校图书馆馆藏建设,无论馆藏数量怎样增长,馆藏各类目占比都较为稳定;高校读者群体更新较快,虽然每年都存在新生入学和毕业生离校的交替,但读者总体数量与身份特征相对稳定。因此,基于稳定数量与分类特征的读者群体,以及稳定基数和占比的馆藏之间,形成大量图书流通数据,为探索读者需求和馆藏流通之间的规律,提供了坚实的数据支撑。

1.1 相关依据

读者作为各自独立且存在差异的社会个体,其行为趋势具有很大的不确定性,采用社会学研究的数理统计方法,对读者图书借阅行为与馆藏流通之间关联关系进行研究,揭示隐藏在随机借阅事件中的规律和变化趋势,并进行描述和预测,对图书馆把握读者知识需求趋势,有针对性地进行图书采访和开展知识服务具有重大的现实意义。相关研究表明,读者借阅动机和使用图书馆的便利性是促成读者发生相关类别图书借阅行为的直接因素;社会分工对读者社会角色的要求,促使读者形成知识需求的动机;读者对其社会角色的自我期待方向及动机强度,推动读者产生知识交流和自我知识建构的图书阅读行为;读者借阅行为是由读者所担当的社会角色以及角色期待,在整体社会文化背景下,结合自身与其他读者之间的差异化特征,在特定的社会知识环境中,在信息与知识交流等因素共同作用下,阅读动机强度短期迅速增强后,发生的知识获取行为与结果。

由读者社会角色和背景构成的读者身份特征,本质上是在社会、文化和知识背景下对读者的分类,是读者借阅相关类别图书的基础。大量研究成果表明,读

* 本文系国家社会科学基金项目"人工智能图书采访决策模型研究"(项目编号:17BTQ026)研究成果之一。
作者简介：王红(ORCID:0000-0003-3418-5181),研究馆员,硕士生导师,E-mail:sxcdwh@163.com;袁小舒(ORCID:0000-0002-1029-6605),硕士研究生;原小玲(ORCID:0000-0002-1957-3736),副研究馆员,硕士;黄建国(ORCID:0000-0002-2424-9615),工程师。
收稿日期:2019-03-26 修回日期:2019-08-20 本文起止页码:59-70 本文责任编辑:王传清

者当前所处的年龄段、社会角色以及读者对未来社会角色和地位的期待是隐藏在读者阅读动机背后最根本、最直接的影响因素。不同类型的读者对某些特定类型的图书具有明显的需求偏好，如女性读者偏爱女性主角的小说[1]，年轻女性侧重爱情文学图书，已婚女性偏爱散文、游记图书[2]；农民工的阅读倾向主要是休闲性文学书籍和实用性较强的技能与考试类书籍[3]。同样，不同类别的图书，尤其是专业性较强的图书，大多指向专业性特征明显的固定类型读者，如古籍文献的主要读者群体，大多是开展科研工作的研究人员[4]。

高校大学生读者除了专业方向不同以外，在微观上还存在地域、家庭教育和个性发展方向等差异特征；而年龄、教育、成长等社会宏观背景，以及对社会、人生、情感和家庭、职业期待的知识需求和储备动机等因素，则具有较强的同质化特征。大学生是具有独特社会角色特征的读者群体，在图书需求方面，除了所学专业相关图书以外，偏爱小说、自然与人文社会科学类的书籍[5]。

由此可见，读者身份特征尤其是读者差异化特征形成的知识需求偏好，能够反映读者图书阅读需求特征。因此，基于相似特征读者的借阅数据，使得描述和预测馆藏流通特征和趋势成为可能。

1.2 问题定义

采用数理分析方法，分析高校本科生读者的图书借阅偏好，需要合理有效地提取与选择读者的特征，根据读者特征设计假设性分析模型，探索不同特征读者与不同分类图书之间的流通关系。

建立合理的推断前提，是开展量化研究的基础。高校图书流通一般具有以下特征：与读者的专业方向有关，专业方向的学习内容通过课程具体表现；与读者的数量有关，包括读者总量和各个专业的读者数量；与读者之间互动有关，主要指联系频繁的读者之间，进行交流互动促成的知识需求；与读者其他需求有关，如确立考研深造目标产生的知识需求，疾病产生的知识需求，以及青年人对情感、婚姻、职业期待方面产生的知识需求等。其中，不同类别图书流通量出现差别的关键因素，最有可能是由读者的专业方向以及读者数量的影响导致的。由于专业分类与图书分类，是两个不同领域的分类，相同名称的分类名词，在概念内涵上，既具有一定知识关联，又分别在各自领域代表不同的意义。因此，还需要通过专业课程的重要载体，如对教材的图书分类，将专业方向和图书分类之间建立起对应关系。读者之间的交流互动，对图书流通产生的影响较为复杂。一方面相同专业、相同入学批次的读者，因学习共同课程，互相交流频繁，共同借阅相关主题类图书，有一些直接规律可循；另一方面，与专业方向无关的图书借阅，则可能缘于读者之间的日常或随机交流内容，此种情况造成的图书流通特点和规律，也需要深入研究。因此，本研究假设读者特征因素是影响图书流通的关键因素，将读者特征因素作为重点考察的变量指标，分析其对分类图书流通量产生影响的程度，由此，本研究将图书流通量与读者特征之间的关系描述为：

读者特征组合 $X=(x_1, x_2, \cdots, x_n)$ 与图书馆的知识分类C具有某种关联和映射关系，根据该类特征的读者与知识互动的流通历史记录作为统计数据，进行相关分析和聚类分析，筛选出特征明显的读者特征，建立回归模型，找到各种因素对图书流通量的影响程度，以及模型对图书流通量预测的拟合优度，从而以严格的数学方法阐释读者特征对图书需求的因果关系，进而探索和揭示隐藏在读者知识需求背后的图书阅读与流通的规律。

定义1：在给定的读者集合R以及读者特征 $x_{ij}=(x_{1j}, x_{2j}, \cdots, x_{ij})$ 中，建立X与馆藏流通 $y_j=(y_1, y_2, \cdots, y_j)$ 之间的函数关系，i表示读者或读者分类，j表示图书或图书分类：

$$y_j = f(x_{1j}, x_{2j}, \cdots, x_{ij}) \qquad 公式(1)$$

本研究的目的在于建立合适的模型，找到合适的读者特征X，利用X解释馆藏流通Y，并能够合理阐释X对Y的因果量化关系，以及根据X预测Y。

1.3 相关研究

以往探究读者借阅与流通内在机制与趋势的研究，主要有以下3种模式。

1.3.1 数据对比模式

基于调查和统计数据，通过量化指标比对，得出读者阅读的倾向与偏好。胡一樱[6]以公共图书馆为例，通过对图书流通统计数据的调查，分析读者阅读倾向及其影响因素，论证了读者阅读倾向分析的必要性。袁红志[7]分析借阅的历史数据，通过藏书流通和图书利用率，了解读者的借阅习惯及读者需求的变化；谢丹玫等[8]利用加权流通率的计算方法进行学生阅读兴趣的主分量分析，在对图书流通清单数据的挖掘和分析基础上，了解学生阅读需求，并改进馆藏结构；周国正和张学敏[9]通过网络问卷调查法，采集在校学生阅读目的、阅读内容、阅读方式以及阅读层次等方面的信

息,并对所获数据进行分析,总结高校学生阅读目的主要为学习与消遣,阅读内容专业特征明显,呈现浅阅读、网络化阅读等特征;吴晓海和黄芳[10]通过对首都医科大学图书馆图书流通数据进行统计分析发现,图书流通量与所学专业和未来工作生活密切相关,与陶冶性情、树立人生观、人际交往等方面息息相关。

1.3.2 相关性分析模式

利用统计数据,假设读者的一些特征与图书流通有关,利用统计学算法进行相关性分析,得到不同因素与读者借阅存在正向或负向相关关系。韩丽[11]通过量表问卷和二阶方程模型,提取并验证了读者的自主动机、基本心理需要满足对读者课外阅读意愿 3 个变量对读者课外阅读意愿产生正向影响,而受控动机对课外阅读意愿则不产生影响;赵雨薇[12]利用关联规则中的 Apriori 算法分析读者需求特征和阅读趋向,为划分读者群细分因素提供合理依据,结合读者自身属性和不同读者的需求特征选取读者细分因素,采取聚类算法细分读者群,建立聚类模型,从而清晰地揭示了读者群需求的差异性;耿倩[13]在对读者的累计借阅数量进行简单贝叶斯分类算法挖掘的过程中,发现图书馆可以通过建立读者个人档案来了解读者背景,挖掘读者的借阅兴趣,进而改进阅读行为,提供更加主动的推荐服务;陈添源[14]利用"网络图"节点和 Apriori 模型进行图书分类号的链接分析和读者借阅图书的分类关联密切程度,并采用关联规则向读者推荐图书。

1.3.3 建模分析模式

通过统计数据,建立各种分析模型,探究读者的行为特征的影响程度,并预测未来的借阅需求趋势。牛秀[15]利用多指数平滑法对华北科技学院图书馆 2007 年 1 月至 2009 年 12 月的月度图书借阅数据进行了实证分析和预测;陈娟和洪丹[16]利用 Logistic 回归模型对用户的借阅影响因素进行分析,结果发现,用户的借阅数量受电子资源、用户课外阅读时间、图书馆环境、用户对课外阅读的认识以及身边好友的影响,而用户的学历、性别、学院特征及主要的借阅动机对用户的借阅数量则没有显著性影响;尹志强[17]针对当前高校图书馆图书借阅流量预测模型存在的精度低难题,引入混沌理论对高校图书馆图书借阅流量原始数据进行分析,建立高校图书馆图书借阅流量建模的学习样本,实验结果表明模型预测图书借阅流量性能更优;张囡和张永梅[18]利用灰色神经网络算法收敛速度较快、误差值小特点,预测图书月借阅量;葛凡[19]采用灰色系统模型和后验差检验准则,预测 TP 类和 TQ 类两个类目图书在未来 5 年的借阅量;田梅[20]采用支持向量机作为建模工具,利用混沌时间序列理论对图书借阅流量行为进行了建模和学习预测;钟亮[21]将用户兴趣爱好、书籍受欢迎程度和用户对书籍的评价等参数组成差异性矩阵,确定各参数影响权重、建模计算用户对书籍的评分,融合采用 k－最近邻分类法和朴素贝叶斯分类法来分类过滤数据,设计实现一种数字图书用户喜好预测算法。

2 研究方法与数据集

2.1 研究方法

馆藏流通的核心问题是解决某类或某种馆藏流通量与读者特征之间的关系问题。很多机器学习的方法都可以完成描述和预测任务,考虑到需要通过研究与苛刻的论证,解释读者特征对馆藏流通的影响关系,本研究采用机器学习中最为成熟和严谨的多元线性回归方法,多元线性回归模型是经济学和其他社会科学中最广泛使用的实证分析工具[22],通过假设某些自变量对因变量具有某种因果作用,建立拟合模型,对模型假设性条件进行全面检验,对具有明显随机性社会事件的馆藏流通量进行科学解释。

馆藏流通与读者身份特征之间的关系可以描述为多元线性回归问题:有 m 个读者或读者类型样本,每个样本对应于 n 维特征和一个流通结果输出 y_m:

$$(x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)}, y_0), (x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)}, y_1), \cdots, (x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)}, y_m)$$

对于 n 维读者特征的样本数据,基于公式(1),构建分类馆藏流通的线性回归拟合模型:

$$y_\theta(x_1, x_2, \cdots, x_n) = \theta_0 + \theta_1 x_1 + \cdots, + \theta_n x_n \qquad \text{公式(2)}$$

公式(2)中 θ_i(i = 0,1,2,…,n)为模型参数,x_i(i = 0,1,2,…,n)为每个样本的 n 个特征值。θ 值就是每个特征对图书流通量的贡献率,θ 的结果采用最小二乘法计算:

$$\theta = (X^T X)^{-1} X^T Y \qquad \text{公式(3)}$$

最小二乘法有着严格清晰的数学推导过程,并可对推导过程进行解释。最终得到包含误差项的样本回归方程:

$$y = \alpha + \hat{\theta} x_i + \hat{\varepsilon}_I \qquad \text{公式(4)}$$

由于可观测数据仅仅是整体读者的抽样样本,最小二乘法是建立在高斯分布基础上,为确保模型的可靠性,还需要对模型的残差进行正态性、残差齐性、残差独立性检验,以及对模型和特征参数进行可靠性检验。

2.2 数据集

数据集来自太原科技大学图书管理系统数据库中 2002 年至 2018 年 6 月的流通数据表、读者信息表、馆藏 MARC 数据以及 2011 年以来各个专业的招生统计数据。数据整合处理后,仅仅保留本科生数据,生成一个包括读者信息、馆藏信息和流通信息等内容的流通记录表,其所包含的字段有读者卡号、读者专业、读者入学时间、图书题名、图书作者、图书分类号、借出时间、借出发生读者的年级。另生成一个招生信息表,包含以下字段:招生时间、专业名称、人数。本研究建立的图书分类表采用《中国图书馆分类法》的分类标准,由英文字母 j 标记不同的图书分类(其中 j = 1,2,…,j),流通数据中的图书分类,根据流通图书的条码,对应该图书的馆藏 MARC 数据中的图书分类号,代表一类图书或一种图书。研究工具采用 R 语言(Version 3.4.1),以及支持线性回归分析的工具包。

数据清洗主要内容是在生成新的流通记录表时,删除数据不完整或缺失关键字段内容的记录,包括通过读者卡号映射读者信息表,因错误不能获取和建立完整的读者信息记录,也包括因为馆藏数据错误,而不能获取和建立完整借阅馆藏信息的记录。最终可用的全部读者流通记录 772 206 条;2011 年以来的读者流通记录 166 763 条,2011 - 2015 年完成 4 年学业的本科生读者流通记录 147 860 条,2015 年以后入学的当前在校本科生流通记录 18 903 条。鉴于图书馆对读者数据收集的限制,以及高校对学生数据管理的具体情况,图书馆能够获取的读者可观测信息极其有限,如该校的读者完整数据存储于学校一卡通系统的数据库中,图书馆的读者数据仅是当读者发生借阅活动时,由图书管理系统向一卡通系统调取读者部分信息数据后,才会将得到的数据存储到图书管理系统中。图书管理系统中仅记录发生借阅行为的读者流通信息,其中图书信息包括题名、条码、借还时间,读者信息包括一卡通号、入学时间、所在院系和专业。

3 研究思路与过程

3.1 读者身份特征提取

根据读者的社会角色来划分,确定研究样本后,对读者进行差异化分组,分组的原则要反映出知识偏好的差异,如果不能分离出组别之间的知识偏好,则分组毫无意义。我国高校本科生读者的基本社会身份是学生身份,绝大多数都是年龄达到 18 周岁,经历基础教育、高中教育后,通过高等教育入学考试,进入高校接受大学本科教育,具有相同或接近的教育内容和知识积累。在入学之后读者最大的差异化特征就是专业方向,专业方向决定读者 4 年的学习内容,也是学生未来人生与职业生涯中极其重要的身份特征。

鉴于数据收集情况,按照可观测特征,将读者集合(R)数据集的读者特征数据分为 3 类:读者专业方向(i_1)、读者专业人数(i_2)、读者入学批次(i_3)。专业方向作为高校学生最基本特征,将读者的类型依据知识学习内容进行初步划分,进一步深入到课程教材的图书分类层面,考察专业分类对专业图书借阅量影响程度;专业人数作为读者特征,可在分类读者数量规模方面,进一步考察读者专业方向的差异性对专业有关图书借阅量的影响程度;专业入学批次作为读者特征,主要考察同一入学批次的读者相互之间的影响与交流密切程度,对图书借阅量的影响程度。

高校读者进行分类的特征还有很多,对一些无法观测的分类特征,如性别特征、入学前居住地特征等,对图书借阅量造成的影响往往会体现在模型的残差中,当残差对模型的精度造成的影响超过模型置信区间要求,则说明研究选择的读者特征无法解释图书流通量。

3.2 读者专业方向分类特征与分类图书流通量关系分析

以读者专业方向作为关键分类特征,探索读者对不同种类图书的偏好,是否存在确定的关联关系,是开展研究的假设性关键前提和基点。利用大数据可视化分析方法,能够简单直观展示出隐藏在数据背后的特征,随机选择 2014 年入学在校 4 年的读者借阅数据,生成流通数据的可视化视图——桑基图(见图 1),图 1 中上层标签代表各个专业借阅总量,底层标签是 22 类图书流通总量,上层和下层的数量相等,中间的连线反映不同专业方向读者和图书分类之间的阅读偏好,连线的宽度反映了不同专业读者和不同图书分类之间的借阅数量。图 1 表明,从专业方向角度看,每个专业方向的读者都有明显的借阅偏好,如借阅量最大的机械设计制造及其自动化专业,T 类图书的借阅量接近该类读者借阅总量的半数;从图书分类角度看,T 类图书的借阅对象主要来自理工科专业方向的读者。其他专业方向的读者也大都具有类似的借阅特征。因此,读者的专业方向与图书分类流通之间具有某种必然的联系。然而由于读者的专业方向并不能直接与图书分类之间建立起必然的联系,还需要深入读者专业方向学习内容层面,寻找更加可靠的证据和关联关系。

王红，袁小舒，原小玲，等.高校图书馆读者借阅趋势线性回归建模预测探析[J].图书情报工作,2020,64(3):59－70.

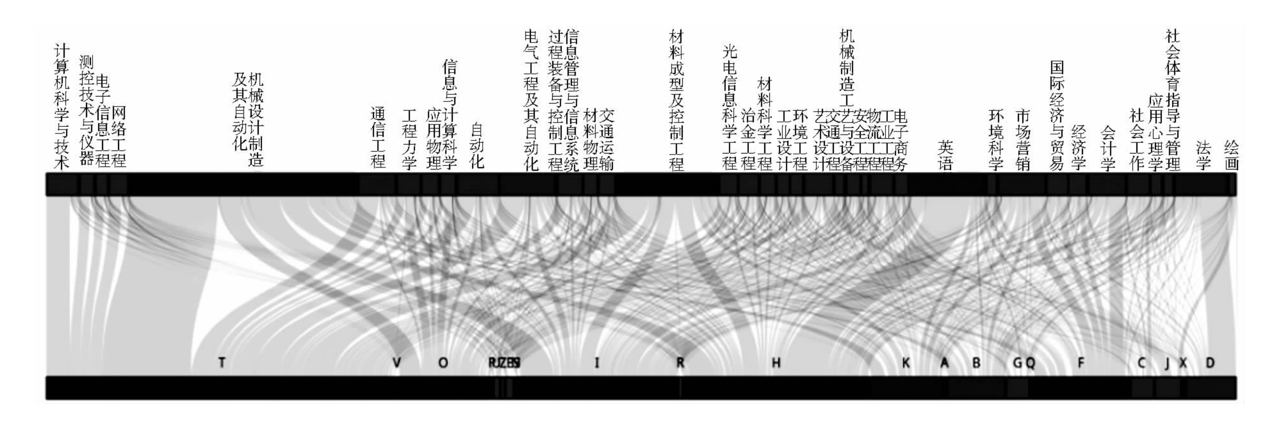


图 1　2014 年入学读者图书分类借阅分布

3.3　专业课程与流通分布趋势分析

课程是专业的具体表达和体现,高校读者主要通过专业课程完成专业学习,考察专业方向与图书分类借阅之间的关系,可利用专业方向作为关键读者特征指标,每个专业方向都是由多门课程组成,相关专业方向的课程设置重复率高。在分析读者专业方向与图书流通之间关系时,深入到专业课程层面,考察图书流通与课程的关系,能够更加准确发现图书流通与专业方向之间的关系。本研究依据该校 57 个专业的《本科专业人才培养方案》,包括通识必修课程、学科基础课程、专业必修课程,将每个专业课程用课程使用教材的图书分类方式进行归类,最后,在图书 22 类层面,汇总每一图书分类中包含的课程数量,得到 22 类总计1 191 门课程,按照图书每一分类的课程数量除以课程总量 1 191门,得到课程分类比。图书流通比的计算方法,是在 22 类图书分类层面,计算每一大类的图书流通数量和该图书馆全部的流通量之比,得到流通比。通过图书馆流通比和课程分类比的对比(见图 2)可以看出,除了 A 类、G 类和 I 类图书以外,课程分类比曲线与流通比曲线变化趋势大体一致,说明专业和图书分类之间呈关联关系。

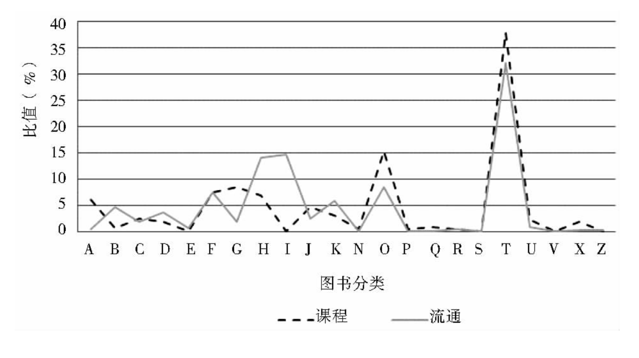


图 2　课程比与流通比对比

3.4　专业课程与图书分类相关性分析

专业方向是由多门课程集合体现,但课程集合的组合,在图书分类上的映射未必唯一,甚至还可能较为分散,因此,关于读者的课程与图书的相关性,还需要做进一步分析验证,才能确保基于课程的专业方向作为研究变量具有采用信度。采用数据可视化方法——热图分析课程设置与图书分类相关性关系见图 3,其中横坐标代表图书分类,纵坐标代表专业方向,中间交叉部分颜色热度代表每个专业课程分类在各个图书分类中的数量大小。通过直接观察发现,各个专业的课程数量的分布,具有明显的差异性。按照各个专业的课程分类包括通识教育课程和核心专业课程,以近年来连续招生的 37 个专业为例,专业课程分类明显具有 4 个层次:第一层次为 T 类课程,专业课程数量最多,课程在各专业之间的分布也更加分散和均匀,呈现明显的工科专业教育特征,其中专业聚类也表明,以机械设计制造及其自动化专业和材料成型及控制工程两个专业的特征最为明显。第二层次为 F、G、O 类课程,在数量上处于第二层次,在课程分布上,O 类与 T 类具有相似的分布特征,呈现分布均匀的状态,根据热图显著性和聚类结果,表明工程力学和材料物理专业,在 O 类课程的数量上更为明显,F、G 类在分布上则集中在少数几个专业上。第三层次范围较大,包括 A、C、D、H、J、K、Q、U、X 等 9 类。A 类和 K 类的分布较为均匀,表明公共通识课的教材分类,主要集中在 A 类和 K 类;其他 7 类图书在分布上较为集中,表明每一类都有能映射到相关专业。第四类包括课程数量较少或没有属于该类的课程。如图 3 中的 B、N、P、R 这 4 类课程较少,I、S、V、Z 这 4 类没有出现课程。

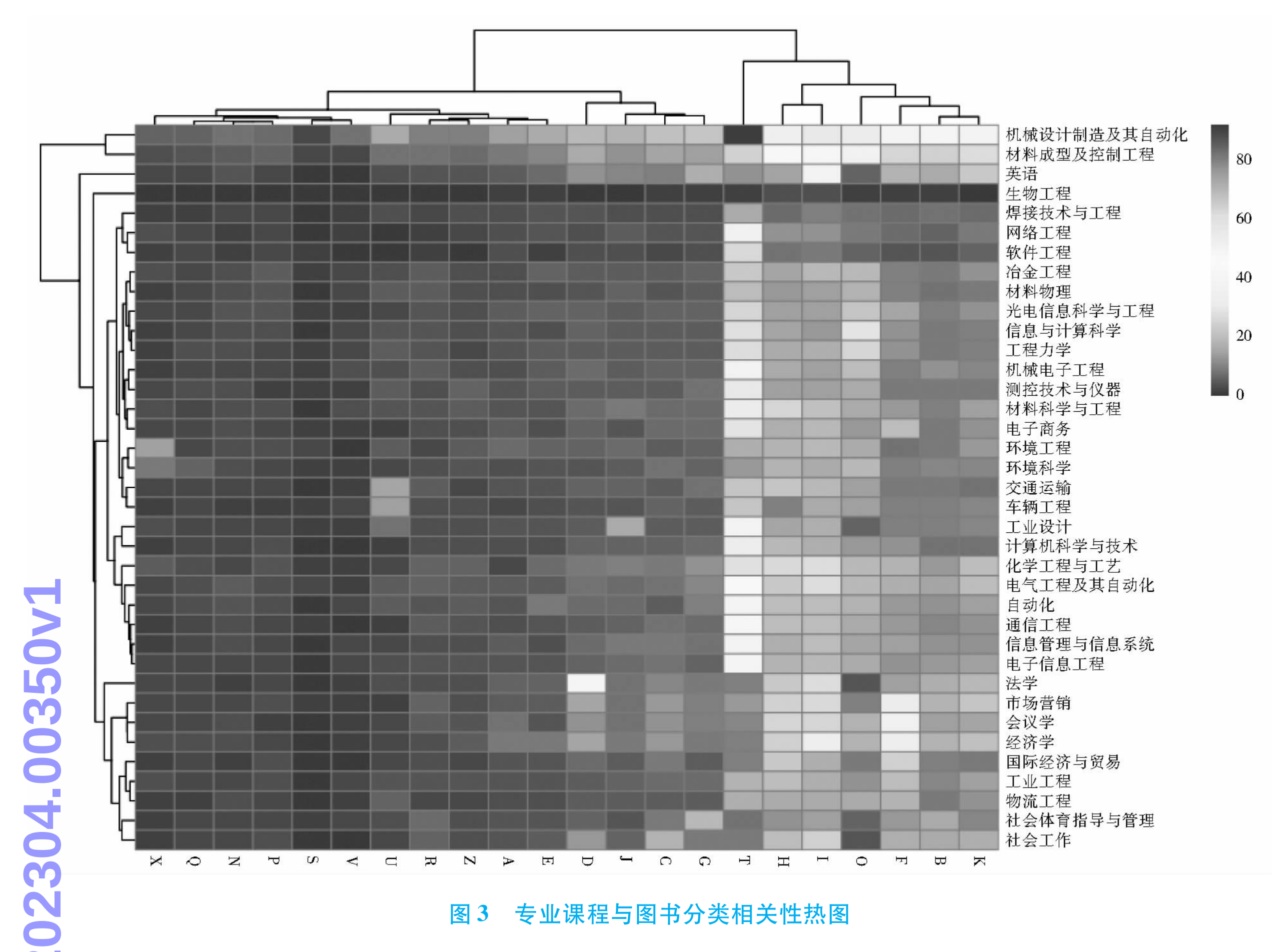


图 3　专业课程与图书分类相关性热图

3.5　专业方向与分类图书流通量相关性分析

专业方向读者的借阅情况,如果能够与专业课程在图书分类上的热点特征吻合,就可以确认无误地把读者专业方向作为建模分析的关键指标。通过统计各个专业方向读者在各个分类的借阅数量,绘制专业图书的大数据流通量分析热图(见图 4),与图 3 相比较而言,分类图书的借阅量情况与图 3 专业课程的借阅情况具有较好的吻合,专业方向读者的借阅偏好更加集中,其中 D 类、F 类、O 类、T 类与图 3 的专业课程分布特征几乎一致,说明本研究基点,即把读者的专业方向作为影响图书流通主要因素的假设成立。同时,图书流通量较大的类别中,I 类并没有任何课程与专业映射,需要关注。

3.6　读者数量与图书流通量的相关性分析

对读者数量和流通量进行 Pearson 相关系数检验,设定置信区间 95%,由表 1 可见,除了 D、G 类图书以外,绝大多数社会科学类图书流通量与读者的数量具有明显的相关性,而自然科学类图书与读者数量没有相关关系,这与人们日常经验不相符。因此,读者数量单一因素并不完全是影响图书流通率的关键因素,还需要将读者数量因素放在不同的分类组合下,经过综合分析后,才能确定读者数量对图书流通量的影响作用。

表 1　读者数量与图书流通量 Pearson 相关系数检验

图书分类	A	B	C	D	E	F	G	H	I	J	K
p 值	0.02	0.00	0.00	0.61	0.00	0.00	0.15	0.00	0.00	0.04	0.00
cor	0.22	0.55	0.36	0.05	0.62	0.25	0.13	0.36	0.65	0.18	0.59

图书分类	N	O	P	Q	R	S	T	U	V	X	Z
p 值	0.00	0.00	0.00	0.02	0.00	0.09	0.00	0.20	0.02	0.65	0.00
cor	0.45	0.63	0.46	0.30	0.47	0.46	0.72	0.15	0.40	-0.07	0.57

3.7　入学批次与图书相关性分析

入学批次是分类变量,采用单因素方差分析方法检验后,得到入学批次对图书流通量的影响。检验结果表明,入学批次仅仅对社会科学中 B、C、F、K 4 类图书流通量有影响。可见单一的专业入学批次因素,对分类图书流通量的影响特征不明显,但专业入学批次因素与其他读者特征组合,对图书借阅量的影响作用,需要在线性回归方法中进一步观察。

4　建模与实验过程

4.1　模型选择

本研究采用最小二乘法进行多元线性回归,基于公式(2),按照图书基本分类的 22 类,分别将每一类的流通量作为因变量,将其他数据作为自变量输入模型,

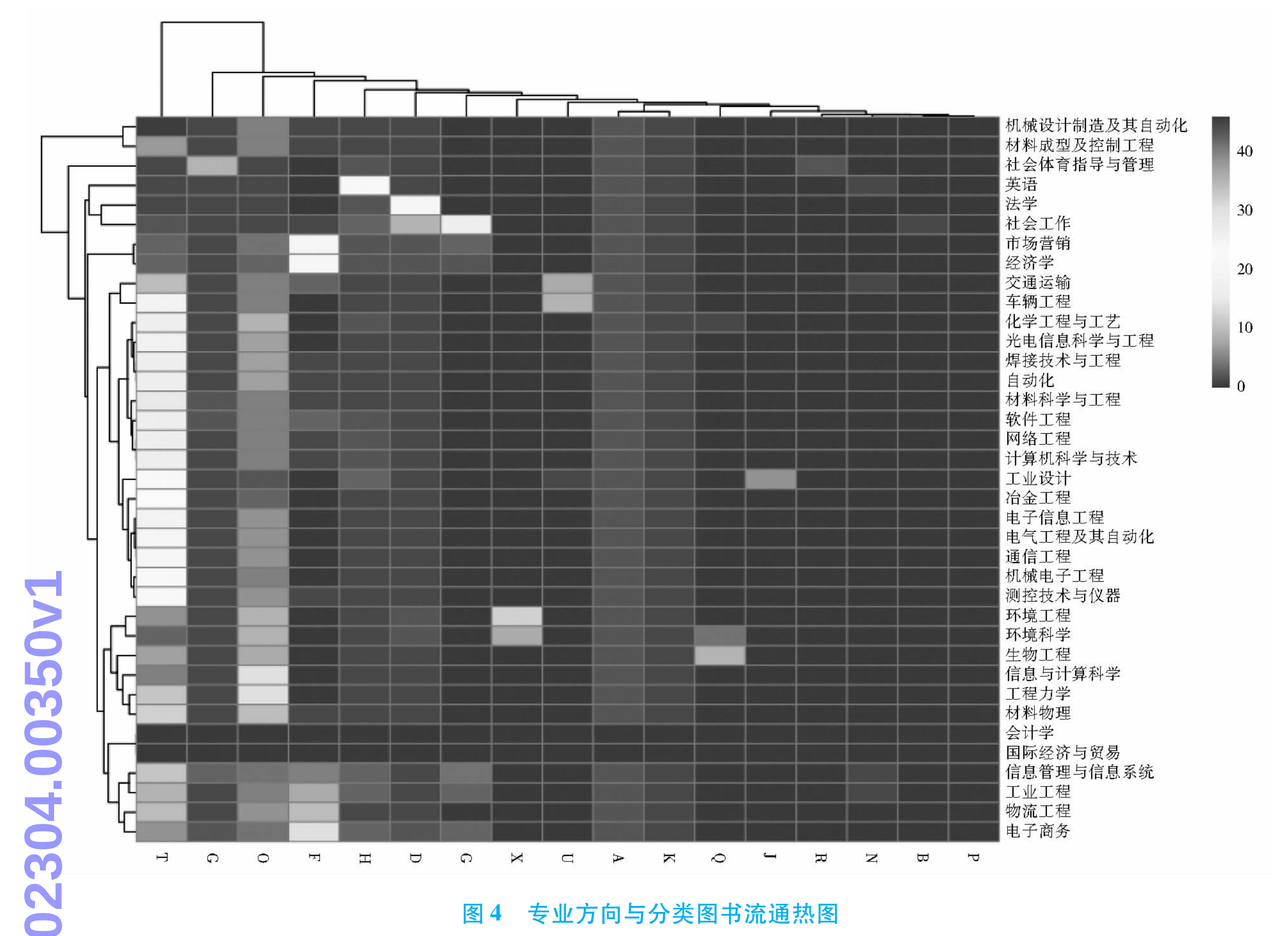


图 4　专业方向与分类图书流通热图

设定 95% 的置信区间，采用逐步法，分别按照 22 类的流通量作为因变量(y_j)、读者专业(x_{1j})、各专业读者数量(x_{2j})、读者入学批次(x_{3j})，导入数据，并对模型优度、模型假设性进行检验。

计算后发现，除了 S 类的数据样本量过小，残差自由度为零，不能建立模型，其他 21 类都可通过模型得到拟合，然而所有的残差都呈指数趋势特征，不符合原始假设，模型没能通过检验。为了保证在分析过程中继续采用线性回归方法，对图书流通量进行对数变换，变量变换后，从 Q-Q 图(见图 5)观察可以初步确定，变量符合线性假设，残差符合正态性要求，除了 S 类图书，其他 21 类图书的流通量拟合模型，均通过检验。

4.2　拟合验证

确立模型后，对模型优度、模型假设性进行检验，结果见表 2。

自然科学部分：总体来看，所有分类的 F 分布的 p 值均小于 0.05，模型有效性通过检验，其中，O、T、U、X 类的模型 R 平方都超过 80%，说明模型解释能力较强。N、P、Q、R 这 4 类图书，从拟合效果看，模型解释能力偏低。

从模型的正态性、同方差检验来看，只有 U 类和 Z 类完全通过检验，其他分类没有通过检验，说明模型的准确率受到其他因素的影响，尤其是在不能引入其他读者特征作为新的变量情况下，需要在因变量的方面寻找突破口。

从变量的系数来看，影响 N、P、Q、R 这 4 类图书流通量，只有读者数量单一特征因素；影响 O、T、U、X 类图书流通量的因素，是专业方向和入学批次两个因素；影响 V、Z 类图书流通量的因素，是读者数量和专业方向两个因素发挥作用。

社会科学部分：模型的拟合值 R 方，都在 50% 以上(见表 3)，拟合效果较好，模型的 F 统计量 p 值都远远小于 0.05，模型有效性通过检验。D-W 检验结果表明，模型有较好的残差独立性，正态性检验结果全部大于 0.05 意味着残差和样本都符合正态分布，同方差检验仅有 E 类模型通过检验，说明其它分类尚存在其他影响因素，与自然科学部分情况相似，也需要在因变量的方面寻找突破口。

4.3　分析与实验——细分因变量

由于在图书的基本分类层面，绝大多数模型的正态性和同方差不能通过检验，意味着样本分布的噪声影响较大，在无法增加实验样本和自变量的情况下，对

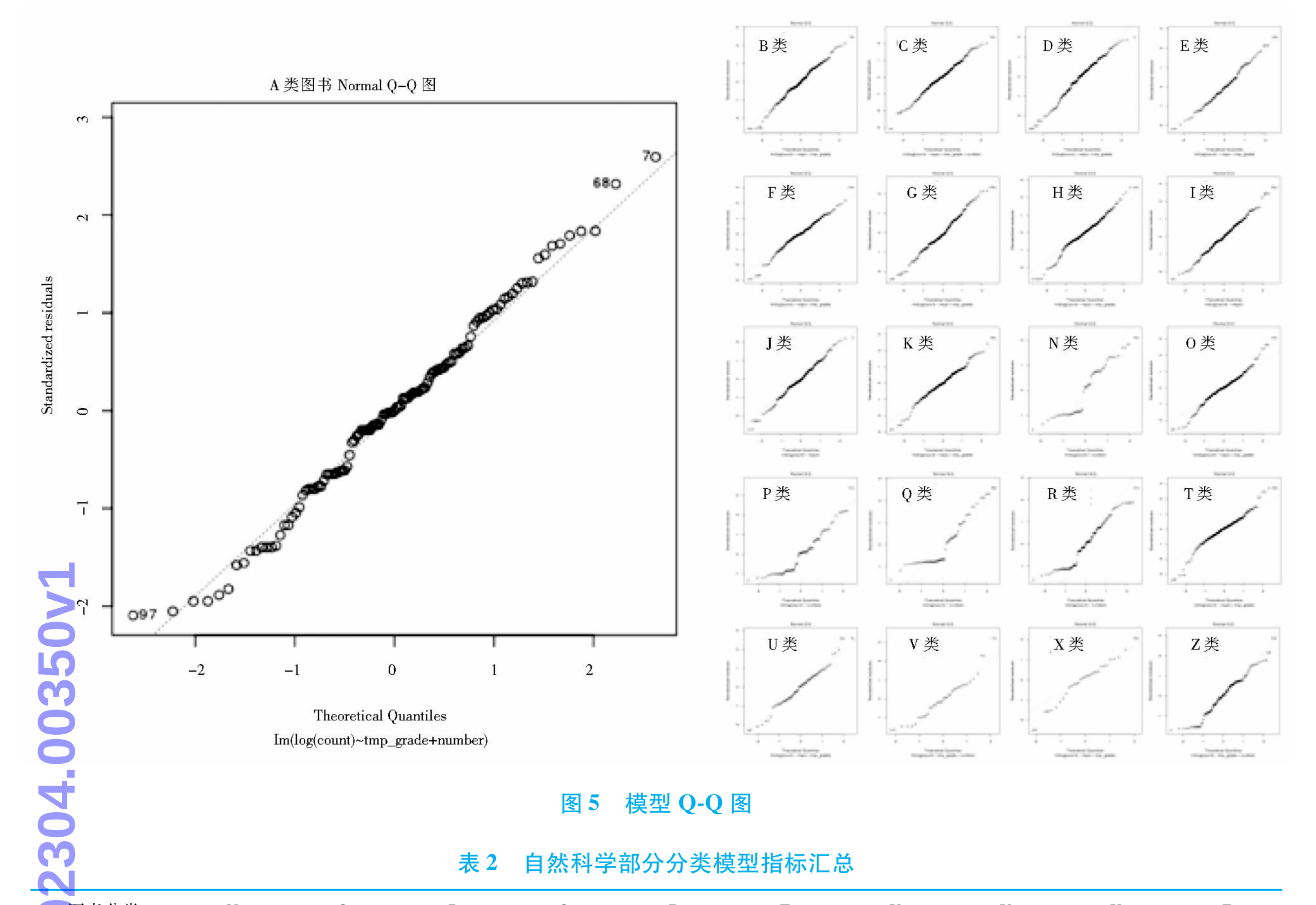


图 5 模型 Q-Q 图

表 2 自然科学部分分类模型指标汇总

图书分类	N	O	P	Q	R	T	U	V	X	Z
R^2	0.129	0.844	0.222	0.096	0.145	0.855	0.83	0.318	0.967	0.539
模型检验 p 值	0.009	0	0	0.018	0	0	0	0.016	0	0.002
同方差检验	0.607	0.007	0.807	0.825	0.927	0.004	0.077	0.047	0.673	0.199
残差独立性检验	2.129	2.311	1.757	1.811	1.898	2.524	2.75	1.852	2.724	2.624
方差正态	0	0.101	0	0	0.001	0.013	0.18	0.101	0.006	0.795
学生数量	1		1	1	1			1		1
入学批次		1				1	1	1	1	1
专业		1				1	1		1	
ncvTest()	0.766	0.131	0.787	0.841	0.945	0	0.68	0.001	0.104	0.888

表 3 社会科学部分分类模型指标汇总

图书分类	A	B	C	D	E	F	G	H	I	J	K
R^2	0.14	0.6	0.69	0.65	0.7	0.64	0.59	0.73	0.77	0.56	0.622
模型检验 p 值	0	0	0	0	0	0	0	0	0	0	0
同方差检验	0.11	0.01	0.08	0.24	0.22	0.01	0.18	0	0	0.05	0.023
残差独立性检验	2.2	2.37	2.15	2.44	2.45	2.28	2.44	2.42	2.48	2.5	2.569
方差正态	0.65	0.65	0.61	0.1	0.65	0.26	0.46	0.11	0.85	0.72	0.293
学生数量	1		1								
入学批次		1	1	1	1	1	1	1			1
专业		1	1	1	1	1	1	1	1	1	1
ncvTest()	0.98	0	0.36	0.18	0.99	0	0.07	0	0	0	0.031

因变量分析还可以考虑一种情况，即图书分类的层次较多，每一分类下属子分类又会对分类进行细化，读者偏好的图书，可能在下一级类目中，被其他子类目干扰。因此，需要更加细致地对图书分类类目进行分析。

在二级分类中，对于模型不能通过检验的 B、F、H、I、J、K、O、T 等 8 类子类进行筛选，找到关键影响的子类，总体来看，从模型的 R^2 来看，对变异的解释能力均超过 50%，最高的 TG 类已经达到 92.8%。

自然科学部分：进入到二级类目后（见表 4），模型的所有检验均得以通过，模型拟合度明显提高，从一级分类中的 O、T 类结果来看，读者的专业特征与图书的流通具有良好的匹配，读者的专业特征也得到良好的反映，值得注意的是，入学批次在绝大多数分类图书流通中被保留下来，这也说明读者之间社会化信息与知识互动对图书流通具有重要影响。

表 4　自然科学部分二级分类图书模型指标

分类	R^2	F 统计量 p 值	Bptest( ) 检验	DW 检验	正态性检验	人数	批次	专业	ncvTest( ) 检验
O2	0.747	0	0.133	2.344	0.977	1	1	1	0.289
O3	0.807	0	0.112	2.333	0.132		1	1	0.675
TB	0.863	0	0.051	2.216	0.259		1	1	0.917
TG	0.928	0	0.151	2.542	0.31		1	1	0.808
TH	0.882	0	0.064	2.682	0.319		1	1	0.384
TM	0.826	0	0.108	2.552	0.052		1	1	0.079
TN	0.894	0	0.143	2.599	0.168		1	1	0.61
TQ	0.766	0.015	0.136	3.019	0.266	1		1	0.02
TU	0.699	0	0.107	2.51	0.743			1	0.507

社会科学部分：社会科学在二级分类中（见表 5）有五大类共 9 个二级子分类通过了模型检验，其中 J 类的 J2、J5 对应的是艺术专业，影响因素是专业特征和入学批次特征。F 类对应的是经济专业，也通过 F8 表现了专业影响力。总体来看，社会科学部分对图书流通起到关键作用的入学批次，读者人数的影响与专业的影响出现 3 次，说明相似年龄特征的读者，社会科学知识的需求具有普遍性，关注热点分别为 B5、I3、K2、K9 类。

表 5　社会科学部分二级分类图书模型指标

分类	R^2	F 统计量 p 值	Bptest 检验	DW 检验	正态性检验	人数	批次	专业	ncvTest 检验
B5	0.617	0	0.211	2.266	0.611		1	1	0.948
F8	0.641	0	0.164	2.68	0.218		1	1	0.237
I3	0.655	0	0.059	2.464	0.234	1	1	1	0.869
J2	0.617	0	0.165	2.692	0.079		1	1	0.205
J5	0.699	0.014	0.387	2.998	0.075		1	1	0.67
K2	0.558	0	0.049	2.829	0.777		1	1	0.302
K9	0.675	0	0.064	2.725	0.55		1	1	0.927

对于未能通过检验的二级分类模型，依然需要进入三级分类、四级分类进行建模分析，直至最后对图书馆全部流通数据进行建模分析（见表 6）。在三级四级图书类目中，有 6 大类 7 个三级子类目、6 个四级子类目通过模型检验。其中，A85 类对应的是读者课程，H31 与公共课程英语对应，F、O、TP2 三大类的子类目与专业对应。I 类文学部分对应的是中国各时期文学的作品集。从相关系数来看，图书流通的关键影响因素依旧是专业方向和入学批次。

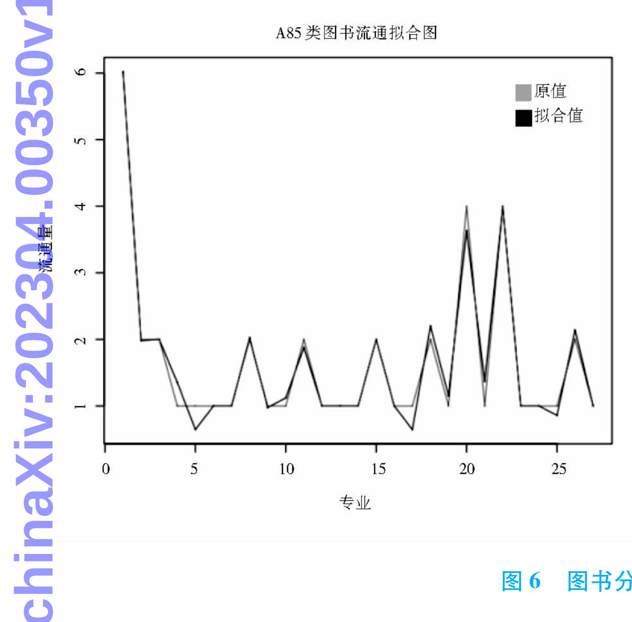
总体来看，读者专业特征能够在较大的图书分类范围内对图书流通量进行解释，读者入学批次和读者数量，需要与读者的专业特征结合，才能具有更加有意义的解释价值。

4.4　拟合和预测

利用模型对原样本数据，以随机选取的 5 个不同层级图书分类为例，对 2014 年不同专业读者流通情况进行模型拟合（见图 6），通过图书流通拟合结果和原值比较，模型拟合值与实值基本相等，并且模型的拟合结果略显保守。这说明选择的变量与图书流通数量之间具有较强的因果关系，能够描述和解释读者图书借阅的需求趋势。

表 6　部分三级分类的模型检验指标

分类	R^2	模型检验 p 值	同方差检验	DW 检验	方差正态	人数	批次	专业	ncvTest
A85	0.978	0.027	0.227	3.014	0.013	1	1	1	0.915
F22	0.739	0.033	0.222	2.441	0.004	1	1	1	0.314
F71	0.598	0	0.194	2.563	0.115		1	1	0.106
H310	0.6	0	0.169	2.504	0.332		1	1	0.811
H311	0.966	0.003	0.183	2.413	0.504		1	1	0.571
H314	0.58	0.001	0.065	2.713	0.226		1	1	0.618
H315	0.645	0	0.002	2.292	0.626		1	1	0.989
I210	0.849	0	0.052	2.647	0.505		1	1	0.697
I217	0.653	0	0.221	2.445	0.342		1	1	0.09
O41	0.826	0	0.122	2.738	0.052		1	1	0.313
O43	0.9827	0.009	281	2.749	0.361	1	1	1	0.127
TP2	0.718	0	0.06	2.289	0.321		1	1	0.178



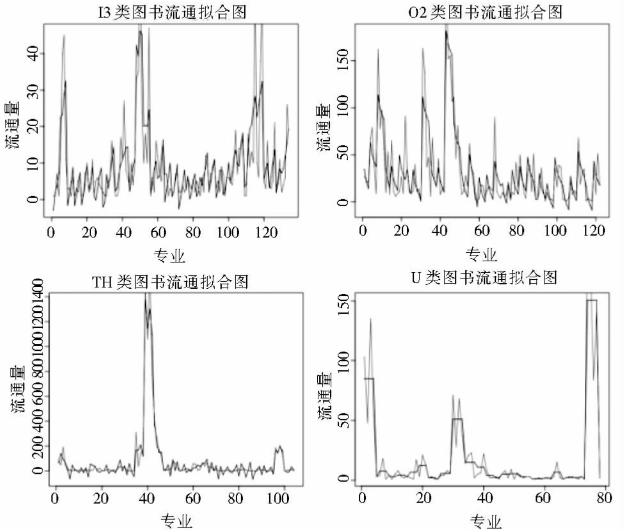


图 6　图书分类模型拟合图

模型对未来图书流通趋势预测是建立模型的关键步骤，模型预测的结果是对某一批次读者总体流通趋势的预测，反映到当前年度的流通趋势，则要在总需求预测中减去不同入学批次读者以及发生的流通结果：

$$P_{ijk} = P_{ij} - C_{ij} \qquad \text{公式(5)}$$

公式(5)中，P 为预测值，C 为已经发生的流通数量，i 为图书分类，j 为入学批次，k 为预测年级，P_{ijk} 为 i 类图书读者未来总流通量预测值，P_{ij} 为批次读者对 i 类图书的总流通预测值，C_{ij} 为 j 入学批次读者 i 类图书已经发生的流通量。

根据公式(5)对当前在校读者(2015 年 – 2018 年入学的学生)，可能发生的图书流通情况进行分类预测，结果见表 7。

从提取出的 20 个图书分类的流通量预测结果来看，对于 2015 年 – 2018 年入学的学生读者借阅趋势，模型具有较好的拟合预测效果，能够对图书流通进行有效描述和准确预测。随着未来读者流通数据量以及可观测数据类型的增加，预测结果的精度将会进一步提高。

5　研究结论

本研究对不同类别图书的流通量因素进行分析，通过相关分析、聚类分析，针对读者集合 R，提取了专业方向、专业方向的读者数量、读者入学时间 3 个有代表性的读者特征因素，作为随机变量，运用线性回归方法，对图书流通进行建模，取得较好的拟合和预测效果。实验结果表明：

(1)图书馆流通可建立数学模型进行描述。看似杂乱的读者随机借阅行为，具有深刻的数理统计规律。通过机器学习的线性回归方法，建立数学拟合模型，能

表 7　模型预测结果

单位:次

预测指标	流预测值(P_{ij})				批次读者实际流通量(C_{ij})				未来总流通量预测值(P_{ijk})			
批次读者	2015	2016	2017	2018	2015	2016	2017	2018	2015	2016	2017	2018
A	165	164	163	160	107	77	3	0	58	87	160	160
B	821	829	831	863	447	140	28	0	374	689	803	863
C	302	309	324	404	133	44	0	0	169	265	324	404
D	500	502	503	513	184	18	9	0	316	484	494	513
E	143	143	142	141	46	27	2	0	97	116	140	141
F	947	955	960	982	280	63	9	0	667	892	951	982
G	239	240	242	252	88	32	15	0	151	208	227	252
H	1 839	1 836	1 833	1 823	1 054	296	75	0	785	1 540	1 758	1 823
I	5 370	5 417	5 457	5 617	3 974	1 662	343	0	1 396	3 755	5 114	5 617
J	381	381	382	382	155	116	31	0	226	265	351	382
K	846	884	901	1012	456	178	59	0	390	706	842	1 012
N	62	62	61	57	20	2	1	0	42	60	60	57
O	2 368	2 427	2 415	2 623	1 574	446	131	0	794	1 981	2 284	2 623
P	90	90	91	92	35	3	1	0	55	87	90	92
Q	89	89	91	94	27	10	4	0	62	79	87	94
R	112	112	112	113	21	6	4	0	91	106	108	113
S	73	72	72	71	2	0	0	0	71	72	72	71
T	6 680	6 766	6 799	7 006	3 118	562	93	0	3 562	6 204	6 706	7 006
U	280	281	280	295	104	37	18	0	176	244	262	295
V	34	34	34	34	7	7	1	0	27	27	33	34
X	97	96	99	98	3	0	1	0	94	96	98	98
Z	105	114	123	160	74	6	3	0	31	108	120	160

够描述分类相同读者的图书借阅行为规律,准确预测图书流通量,也能够合理解释读者借阅行为的内在动机因素和外在的社会交流因素。

(2)知识需求是读者分类的关键。读者的知识行为驱动,虽然与读者自身的修养、工作和生活息息相关,与读者社会关系角色的工作生活内容密切相连。但从研究的结果来看,代表读者知识需求的重要特征——读者的专业学习方向,与特定的图书分类具有直接关联关系,并在模型中起到关键作用;而专业读者的数量和读者入学批次等知识需求特征不明显的因素,对图书流通的影响作用并不明显。对于没有观测到的读者特征,如读者年龄阶段下有关社会、情感、婚姻等知识需求的特征,分别在相关分类图书和模型残差中得以体现。

(3)自然科学类图书读者易于划分识别边界。从图书分类角度来看,自然科学与理工类流通图书的读者行为与特征相对易于描述分析,说明理工类知识的专业性较强,非专业读者涉猎较少。利用读者的专业特征,就能很好地对读者进行分类,分类后的读者群体专业特征相近,图书借阅的特征指向较为清晰,模型拟合效果较好。

(4)社会科学类读者区分边界模糊。社会科学方面的图书流通,专业类的读者和非专业类读者混杂度较高,模型的拟合度往往不高。这说明读者作为社会化的个体,汲取社会知识的内在动力因素更加复杂,仅仅依靠读者专业方向尚不能较好对读者分类进行分隔,还需要挖掘更多的读者细分特征,进行更加深入的研究,才能提高模型拟合预测的精度,发现更加隐秘的规律。

6　结语

本研究运用线性回归分析方法,以易于获取的读者特征对读者进行分类,以高校本科生读者专业方向、专业方向读者数量和读者入学批次 3 个量化指标作为建模分析的关键变量,描述读者需求并预测图书馆流通趋势。不仅为揭示读者借阅行为提供了方法与借鉴,同时也为采用读者分类特征进行知识获取行为分析,提供了探索方向和研究思路。通过分析读者借阅的内在心理动机,为进一步挖掘读者产生知识需求的促动因素,探讨图书借阅行为发生的动机强度等提供了一个可能的突破口。由于获取数据单一,为使读者分类特征稳定,本文仅选择了一个学校的图书馆读者群体作为样本,在后续研究中,将不断获取新的数据对模型进行验证,以期使研究扩展到更多类型的高校图书馆、公共图书馆,使研究结论具有更广泛实用意义。

参考文献:

[1] GIDDENS A. Sociology[M]. Cambridge:Polity Press. 2009.
[2] SUMMERS K. Adult reading habits and preferences in relation to gender differences[J]. Reference & user services quarterly, 2013,

52(3):243 – 249.
[3] 周甜甜. 女性阅读与图书馆服务探微[J]. 大学图书情报学刊, 2014(3):105 – 108.
[4] 舒曼. 新生代农民工阅读倾向与成就动机、心理控制源关系研究[J]. 中国出版,2017(24):29 – 33.
[5] 薛文辉. 古籍读者阅读倾向调查分析[J]. 图书馆学刊,2018(3):89 – 94.
[6] 胡一樱. 公共图书馆读者阅读倾向实例研究——以绍兴市柯桥区图书馆为例[J]. 图书馆研究与工作,2015(2):14 – 17.
[7] 袁红志. 从图书流通数据透视馆藏结构及读者阅读倾向——以衡阳师范学院图书馆为例[J]. 衡阳师范学院学报,2016(2):170 – 173.
[8] 谢丹玫,徐荣华,陆飞,等. 大学生的阅读需求与馆藏建设[J]. 情报探索,2014(2):68 – 72,75.
[9] 周国正,张学敏. 大学生阅读倾向对高校图书馆利用的影响[J]. 情报探索,2016(8):84 – 86.
[10] 吴晓海,黄芳. 首都医科大学医学生图书借阅行为分析[J]. 中华医学图书情报杂志,2015(5):44 – 49.
[11] 韩丽. 自我决定理论视角下高校读者阅读意愿影响因素探究[J]. 图书情报工作,2018,62(14):22 – 27.
[12] 赵雨薇. 基于数据挖掘感知读者需求的高校图书馆差异化服务研究[J]. 图书馆工作与研究,2018(7):68 – 73.
[13] 耿倩. 贝叶斯算法在图书馆读者智能分析中的应用[J]. 自动化技术与应用,2018(5):14 – 16.
[14] 陈添源. 高校读者借阅行为的关联分析及应用实践[J]. 情报探索,2018(12):97 – 102.
[15] 牛秀. 基于多参数指数平滑的图书借阅量预测[J]. 科技情报开发与经济,2011(28):50 – 51.
[16] 陈娟,洪丹. 基于 Logistic 模型的高校图书馆用户借阅影响因素分析[J]. 情报科学,2013(3):96 – 101.
[17] 尹志强. 基于数据挖掘的高校图书馆图书借阅流量建模与分析[J]. 微电子学与计算机,2018(11):95 – 99.
[18] 张囡,张永梅. 基于灰色神经网络的图书馆图书借阅量预测[J]. 情报探索,2013(3):133 – 135.
[19] 葛凡. 基于灰色系统模型的图书借阅量预测分析[J]. 教育教学论坛,2018(11):106 – 109.
[20] 田梅. 基于混沌时间序列模型的图书借阅流量预测研究[J]. 图书馆理论与实践,2013(7):1 – 3,26.
[21] 钟亮. 用 k – 最近邻和贝叶斯分类预测图书用户喜好[J]. 信息技术,2016(9):62 – 65.
[22] WOOLDRIDGE J. Introductory econometrics[M]. Mason:Cengage Learning,2009.

作者贡献说明:
王红:论文构思与撰写,模型实验,论文修改;
袁小舒:数据清洗,数据相关性分析,数据集预处理;
原小玲:数据采集,数据统计;
黄建国:R 语言编程,模型校验。

Prediction of Reader Lending Trend in Academic Library by Linear Regression Modeling

Wang Hong[1] Yuan Xiaoshu[2] Yuan Xiaoling[3] Huang Jianguo[4]
[1] Library, Shanxi University of Finance and Economics, Taiyuan 030006
[2] School of Information, Shanxi University of Finance and Economics, Taiyuan 030006
[3] Library, Taiyuan Science and Technology University, Taiyuan 030024
[4] Taiyuan Daran Science and Technology Co. Ltd, Taiyuan 030006

Abstract: [**Purpose/significance**] By means of the classification and circulation data of library collection, the paper finds the close correlation between reader characteristics and library collection circulation, establish the relationship model. And through model fitting and prediction, this study explores the implicit rule between reader and library circulation which provides technical and means support for the intelligent management of library. [**Method/process**] Firstly, this paper used clustering and correlation analysis techniques to extract the macroscopic observable characteristics of readers, constructed the direct and indirect mapping relationship between reader characteristics and book classification, and then constructed the regression model of the circulation of reader characteristics and classified books, and verified the validity of the model and optimized the goodness of fit of the model. According to the effective model, this paper explored the trend change of library circulation, and sum up the underlying rules of knowledge construction of the macroscopic characteristics of readers, as well as the impact on the circulation of books. [**Result/conclusion**] There are 3 classification characteristics of readers, namely, the professional learning direction representing the social role requirements of readers, the enrollment batch representing the interaction effect between readers and the number of readers, which can effectively fit and predict the book circulation. The prediction results show that the model has high accuracy and can be used as an effective tool to provide reliable technical support for library to develop knowledge service.

Keywords: university libraries circulation prediction data mining linear regression